Department for Applied Statistics
Johannes Kepler University Linz

IFAS Research Paper Series
2006-19

Randomisierte Befragungsdesigns bei Ziehen ohne Zurücklegen

Andreas Quatember

December 2006

Zusammenfassung

Randomisierte Antworttechniken bieten bei unangenehmen Themen eine alternative Möglichkeit zur Erzielung unverzerrter Schätzer für interessierende Parameter, wenn die direkte Befragung das Risiko hoher Nonresponse- bzw. Falschantwortraten birgt. Man bezahlt dafür mit einer Erhöhung der Ungenauigkeit der Schätzer. In diesem Aufsatz werden für zwei solche Schätzer für Anteile die Varianzen bei Ziehen ohne Zurücklegen hergeleitet und die Qualität verschiedener solcher Schätzer, ihre Unterschiede und ihre Grenzen betrachtet.

1 Einleitung

Der Bereich der statistischen Stichprobentheorie setzt sich mit den Auswirkungen unterschiedlicher Stichprobenverfahren und verschiedener Schätzmethoden auf die Qualität von aus Stichproben gewonnenen Schätzern für unbekannte Parameter auseinander. Sie ist eine reine Fullresponsetheorie, die in ihrem Urnenmodell den Fall nicht vorsieht, dass ausgewählte Kugeln sich nicht in der Urne befinden oder uns die Auskunft über ihre Farbe verweigern. Mit diesem Problemfeld beschäftigt sich die Theorie der fehlenden Werte, deren prominenteste Protagonisten die Amerikaner Roderick J.A. Little und Donald B. Rubin sind. Ein Ziel des am IFAS - *Institut für Angewandte Statistik* der Johannes Kepler Universität Linz eingerichteten institutsinternen Forschungsprojekts „Datenqualität in Stichprobenerhebungen“ ist es, ein in der Praxis anwendbares Manual zu erstellen, das diese beiden Theorien miteinander verknüpft (für nähere Informationen zu diesem Forschungsprojekt siehe in der Forschungsdokumentation der Johannes Kepler Universität Linz: `https://fodok.jku.at/fodok/forschungsprojekt.xsql?FP_=1422`).

Das Auftreten von Antwortausfällen, also von „Nonresponse“ in Stichprobenerhebungen kann sich sowohl in einer verzerrten Schätzung von Parametern als auch in einer Erhöhung der Ungenauigkeit der erhobenen Schätzer niederschlagen. Bevor statistische Methoden wie die Gewichtsanpassung und die Imputation zur Kompensation aufgetretenen Nonresponses eingesetzt werden, ist natürlich dafür Sorge zu tragen, diesen so gering wie möglich zu halten. Denn selbst im besten Fall lässt sich auf diese Weise zumindest eine Erhöhung der Varianz des Schätzers im Vergleich zum Fullresponsefall nicht vermeiden. Sofern eine Ursache für den Antwortausfall die Verweigerung der Beantwortung auf Grund einer „heiklen“ Thematik (wie Drogenmissbrauch, Sexualverhalten, Abtreibungen oder kriminelle Delikte) ist, kann die Verwendung von für solche Fälle entwickelten *randomisierten Antworttechniken* dazu beitragen, die Verweigerungsrate zu minimieren. Den Befragungsstrategien, die unter diesem Begriff subsumiert werden, ist gemeinsam, dass es dem Interviewer durch Anwendung eines vorgegebenen Befragungsdesigns unmöglich gemacht wird, die konkrete Frage, auf die der Respondent tatsächlich antwortet, zu identifizieren. Auf diese Weise kann der Interviewer aus der gegebenen Antwort nicht auf das Vorliegen bzw. Nichtvorliegen der heiklen Eigenschaft schließen. Die Idee dabei ist, dass dadurch dem Interviewten die Angst vor einem ihm unangenehmen „Outing“ gegenüber dem Interviewer so weit genommen werden kann, dass er – unabhängig davon, welche Frage er tatsächlich zur Beantwortung erhält – diese wahrheitsgetreu beantwortet bzw. die Antwort darauf nicht überhaupt verweigert.

2 Methodenüberblick

2.1 Die grundlegenden Verfahren

Als die Pionierarbeit auf dem Gebiet der randomisierten Antworttechniken darf Warners Aufsatz aus dem Jahre 1965 betrachtet werden. Seien U die Grundgesamtheit von N Erhebungseinheiten, U_A die Menge jener N_A Elemente aus U, die eine sensitive Eigenschaft A aufweisen und U_{A^c} die Menge der N_{A^c}-elementigen Gruppe, die diese nicht aufweisen ($N = N_A + N_{A^c}$, $U = U_A \cup U_{A^c}$). Der interessierende Parameter π_A sei die relative Größe von U_A:

$$\pi_A = \frac{N_A}{N} \tag{1}$$

Warners Befragungsdesign lässt sich dadurch charakterisieren, dass der Interviewte (z.B. durch zufällige Ziehung einer Karte) entweder mit Wahrscheinlichkeit p die Frage „Gehören Sie zur Gruppe U_A?“ oder mit Wahrscheinlichkeit $1-p$ die alternative Frage „Gehören Sie zur Gruppe U_{A^C}?“ zu beantworten hat. Sei $\pi_A^{(i)}$ der Anteil der Träger der Eigenschaft A in der „Urne“ zum Zeitpunkt der Befragung der i-ten Erhebungseinheit der Stichprobe s, so ist die Wahrscheinlichkeit $\pi_y^{(i)}$ für eine „ja-Antwort“ des Respondenten i:

$$\pi_y^{(i)} = p \cdot \pi_A^{(i)} + (1-p) \cdot (1 - \pi_A^{(i)})$$

Bei Ziehen der Erhebungseinheiten uneingeschränkt zufällig mit Zurücklegen aus U ist $\pi_A^{(i)} = \pi_A \; \forall \, i \in s$ und mit $\hat{\pi}_y$, dem Anteil der „ja-Antworten“ in der Stichprobe, ist damit

$$\hat{\pi}_A^W = \frac{\hat{\pi}_y + p - 1}{2p - 1} \tag{2}$$

(mit $p \neq 0,5$) unverzerrter Schätzer für π_A (1) (vgl. Warner (1965), S.65). Da bei Ziehen ohne Zurücklegen gilt: $E(\pi_A^{(i)}) = \pi_A \; \forall \, i \in s$, gilt diese Schätzereigenschaft von $\hat{\pi}_A^W$ auch bei diesem Stichprobenverfahren. Die Varianz von $\hat{\pi}_A^W$ ist unabhängig vom verwendeten Stichprobenverfahren eine Funktion der Varianz des Anteils an „ja-Antworten“ in der Stichprobe: Bezeichnen wir mit y_i die Antwort der i-ten Erhebungseinheit und ist

$$y_i = \begin{cases} 1 & \text{wenn i mit „ja“ antwortet,} \\ 0 & \text{sonst.} \end{cases}$$

Dann gilt mit $\widehat{\pi}_y = \frac{\sum_s y_i}{n}$:

$$\begin{aligned} V(\hat{\pi}_A^W) &= \frac{1}{n^2 \cdot (2p-1)^2} \cdot V(\sum\nolimits_s y_i) \\ &= \frac{1}{n^2 \cdot (2p-1)^2} \cdot [E(\sum\nolimits_s y_i^2) + E(\sum\nolimits_{s(i \neq j)} y_i \cdot y_j) - E^2(\sum\nolimits_s y_i)] \end{aligned} \tag{3}$$

Es ist der Wert des Summanden in der Mitte der eckigen Klammer von (3), der sich bei uneingeschränkter Zufallsauswahl ohne Zurücklegen von jenem bei Ziehen mit Zurücklegen unterscheidet. Denn bei Ziehen ohne Zurücklegen bleiben die Wahrscheinlichkeiten dafür, dass sich ein Respondent in U_A befindet, im Gegensatz zu Ziehen mit Zurücklegen nicht konstant. Dies führt zu einer deutlichen Erhöhung des Aufwands bei der Herleitung der Schätzervarianz. Gemeinsam mit der ebenso erhöhten Komplexität in ihrer formalen Darstellung liegen somit die Gründe dafür vor, dass in nahezu allen Veröffentlichungen über randomisierte Antworttechniken ausschließlich Ziehen mit Zurücklegen betrachtet wird. In kleinen Grundgesamtheiten (wie Betrieben, in denen z.B. der Anteil an Beschäftigten geschätzt werden soll, die Büromaterial entwenden) führt eine Verwendung des so berechneten Varianzschätzers aber bei tatsächlichem Ziehen der Personen ohne Zurücklegen zu einer Überschätzung der Ungenauigkeit von Schätzern wie $\hat{\pi}_A^W$, die sich in Form zu breiter approximativer Konfidenzintervalle manifestiert.

Die Schätzervarianz (3) ist bei Warners Technik bei uneingeschränkter Zufallsauswahl von n Elementen und Ziehen mit Zurücklegen (bzw. bei Ziehen ohne Zurücklegen aus einer großen Grundgesamtheit annähernd)

$$V(\hat{\pi}_A^W) = \frac{\pi_A \cdot (1 - \pi_A)}{n} + \frac{p \cdot (1-p)}{n \cdot (2p-1)^2} \tag{4}$$

(siehe: Warner (1965), S.65). Nach einiger Rechnung gilt dafür bei Ziehen ohne Zurücklegen (vgl. Kim and Flueck (1978), S.347):

$$V(\hat{\pi}_A^W) = \frac{\pi_A \cdot (1 - \pi_A)}{n} \cdot \frac{N-n}{N-1} + \frac{p \cdot (1-p)}{n \cdot (2p-1)^2} \tag{5}$$

Zur unverzerrten Schätzung von $V(\hat{\pi}_A^W)$ ist in (4) bzw. (5) π_A durch $\hat{\pi}_A^W$ und im Nenner des ersten Summanden n durch $n-1$ zu ersetzen.

Der jeweils rechte der beiden Summanden in (4) und (5) ist die Varianzzunahme, die durch Warners Befragungsdesign im Vergleich zur direkten Befragung verursacht wird. Die Varianzen $V(\hat{\pi}_A^W)$ sind umso größer, je näher sich die Wahrscheinlichkeit p an 0,5 befindet. Gleichzeitig jedoch garantiert gerade eine solche Wahrscheinlichkeit ein erhöhtes Maß an Vertrauen in die Wahrung der Privatsphäre des Respondierenden. Diesen gegenpoligen Interessen muss bei der Wahl von p Rechnung getragen werden. Dieser Designparameter sollte demnach idealerweise so festgelegt werden, dass er so weit von $0,5$ entfernt ist, dass gerade noch die volle Kooperationsbereitschaft gewährleistet wird.

Ist $p = 1$ (bzw. $p = 0$), so wird direkt nach der Zugehörigkeit zu U_A gefragt. Die direkte Befragung ist somit jener Sonderfall von Warners Befragungsdesign, der die maximale Respondentenbelastung durch den geringsten Schutz der Privatsphäre mit der daraus resultierenden größten Nonresponse- bzw. Falschantwortwahrscheinlichkeit darstellt.

Greenberg et al. (1969) führten eine Idee, die Horvitz et al. (1967) präsentierten, theoretisch aus und trugen damit das zweite grundlegende Verfahren der randomisierten Antworttechniken bei. Ihr Befragungsdesign unterschied sich von jenem Warners durch das Ersetzen der alternativen Frage nach der Zugehörigkeit zur Menge U_{A^c} durch die Frage nach der (mit der Zugehörigkeit zu U_A nicht zusammenhängenden) Zugehörigkeit zu einer Gruppe U_B der Größe N_B. Elemente der Menge U_B sollen sich als Träger einer Eigenschaft B auszeichnen, die als unverfänglich empfunden wird (z.B. der Geburtsmonat B oder das Wohnbundesland B). Mit $\pi_B = \frac{N_B}{N}$, ist

$$\hat{\pi}_A^G = \frac{\hat{\pi}_y - \pi_B \cdot (1-p)}{p} \tag{6}$$

unverzerrter Schätzer von (1). Die Varianz des Schätzers (6) ist bei Ziehen mit Zurücklegen gegeben durch

$$V(\hat{\pi}_A^G) = \frac{\pi_y \cdot (1 - \pi_y)}{n \cdot p^2} \tag{7}$$

(vgl. Greenberg et al. (1969), S.533). Diese Varianz lässt sich durch Ersetzen von π_y, der Wahrscheinlichkeit für eine „ja-Antwort“, durch $\hat{\pi}_y$ und n durch $n-1$ erwartungstreu schätzen. Bei Ziehen ohne Zurücklegen gilt:

$$V(\hat{\pi}_A^G) = \frac{\pi_y \cdot (1-\pi_y)}{n \cdot p^2} - \frac{n-1}{n \cdot (N-1)} \cdot [\pi_A \cdot (1-\pi_A) + (\frac{1-p}{p})^2 \cdot \pi_B \cdot (1-\pi_B)] \quad (8)$$

Der **Beweis** dafür befindet sich im Anhang. Der (positive) Subtrahend auf der rechten Seite von (8) ist der Effizienzgewinn durch Ziehen der Erhebungseinheiten ohne Zurücklegen im Vergleich zur mit Zurücklegen durchgeführten Ziehung. In großen Grundgesamtheiten gilt: $(7) \approx (8)$.

Ist π_B unbekannt, so ist für dieses Verfahren eine modifizierte Vorgehensweise zu wählen, um auch diese Variable schätzen zu können (siehe ebd., S.523ff). Hierbei wird die oben geschilderte Strategie in zwei unabhängigen Stichproben s_1 und s_2 mit Umfängen n_1 und n_2 $(n_1 + n_2 = n)$ angewendet. Die Wahrscheinlichkeiten in den beiden Stichproben für die Frage nach der Zugehörigkeit zu U_A werden mit p_1 und p_2 bezeichnet und es hat zu gelten: $p_1 \neq p_2$. Der daraus resultierende erwartungstreue Schätzer $\hat{\pi}_A^{G'}$ für π_A ist (siehe ebd., S.525):

$$\hat{\pi}_A^{G'} = \frac{\hat{\pi}_{y_1} \cdot (1-p_2) - \hat{\pi}_{y_2} \cdot (1-p_1)}{p_1 - p_2} \quad (9)$$

Greenberg et al. (1969) entwickelten die Varianz des Schätzers (9) bei Ziehen mit Zurücklegen (siehe ebd., S.525). Die überaus komplexe Varianzformel für Ziehen ohne Zurücklegen lässt sich bei Kim and Flueck (1978) nachschlagen (ebd., S.348). Diese Varianzen hängen wesentlich von der Wahl der Wahrscheinlichkeiten p_1 und p_2 für das Stellen der heiklen Frage, der Stichprobenumfänge n_1 und n_2 und des Anteils π_B ab. Greenberg et al. (1969) wiesen nach, dass p_1 und p_2 so weit wie möglich auseinanderliegen sollen und sie schlugen als Faustregel $p_1 + p_2 = 1$ vor. Ferner hat hinsichtlich der optimalen Allokation des Gesamtstichprobenumfangs auf die beiden Stichproben zu gelten (vgl. ebd., S.528):

$$\frac{n_1}{n_2} = \sqrt{\frac{\pi_{y_1} \cdot (1-\pi_{y_1}) \cdot (1-p_2)^2}{\pi_{y_2} \cdot (1-\pi_{y_2}) \cdot (1-p_1)^2}}$$

Da die Wahrscheinlichkeiten π_{y_1} und π_{y_2} für eine „ja-Antwort“ in s_1 bzw. s_2 auch von π_A und π_B abhängen, müssen diese zu diesem Zweck (z.B. aus früheren Erhebungen) – zumindest grob – geschätzt werden. Der größere Teil von n wird dann jener Stichprobe zugeteilt, für die p_i größer gewählt wurde $(i = 1, 2)$. Das Merkmal B wiederum sollte danach ausgewählt werden, dass der Anteil π_B auf derselben Seite von 0,5 wie π_A so weit wie möglich von 0,5 entfernt ist (siehe ebd., S.526ff). Moors (1971) ergänzte die Überlegungen zur Festlegung der Wahrscheinlichkeiten p_1 und p_2 dahingehend, dass konsequenterweise für $p_1 > 0,5$ die Wahrscheinlichkeit $p_2 = 0$ gewählt werden und die 2. Stichprobe somit zur Gänze der Schätzung von π_B dienen sollte.

2.2 Modifikationen der grundlegenden Verfahren

Warners und Greenberg et al.’s Befragungsdesigns bildeten die Basis für Weiterentwicklungen im Bereich randomisierter Antworttechniken. Eine Gruppe solcher Entwicklungen, die wir für unseren Vergleich unterschiedlicher solcher Verfahren heranziehen wollen, bilden zweistufige Befragungsdesigns. Diese sind dadurch charakterisierbar, dass sie vor Warners bzw. Greenberg et al.'s Methoden eine zusätzliche Randomisierungsstufe einschieben. Mangat and Singh (1990) etwa stellen in der 1. Stufe mit Wahrscheinlichkeit q die Frage „Gehören Sie zur Gruppe U_A?", während der Respondent mit der Gegenwahrscheinlichkeit $1-q$ zu Warners Vorgehensweise mit dem Designparameter p verwiesen wird. Der erwartungstreue Schätzer $\widehat{\pi}_A^{MS}$ für π_A ist gegeben durch (siehe ebd., S.440):

$$\hat{\pi}_A^{MS} = \frac{\hat{\pi}_y - (1-q)\cdot(1-p)}{p + (2q-1)\cdot(1-p)} \tag{10}$$

In einem weiteren zweistufigen Befragungsdesign (Mangat (1992)) wird in der 1. Stufe mit Wahrscheinlichkeit q die Frage „Gehören Sie zur Gruppe U_A?" gezogen, während mit Wahrscheinlichkeit $1-q$ auf Greenberg et al.'s Design bei bekanntem π_B verwiesen wird. Der unverzerrte Schätzer $\widehat{\pi}_A^{M1}$ für π_A zu diesem Befragungsdesign ist mit dem Designparameter p gegeben durch (siehe ebd., S.84):

$$\hat{\pi}_A^{M1} = \frac{\hat{\pi}_y - (1-q)\cdot(1-p)\cdot\pi_B}{p + q\cdot(1-p)} \tag{11}$$

Die Varianzen der Anteilschätzer $\widehat{\pi}_A^{MS}$ und $\widehat{\pi}_A^{M1}$ ergeben sich aus den Varianzen der Schätzer $\hat{\pi}_A^W$ und $\hat{\pi}_A^G$, wenn darin die Wahrscheinlichkeit p durch $q+(1-q)\cdot p$ ersetzt wird. Die Vergleichbarkeit dieser Varianzen ist natürlich nur solange gegeben, solange die Befragten durch die zweistufige „Verschleierung" dieser Erhöhung der Wahrscheinlichkeit für die Frage nach U_A tatsächlich keinen Verlust an Privatsphäre empfinden und sich deshalb kooperationswillig zeigen. Auch die Wahrscheinlichkeit $q+(1-q)\cdot p$ darf demnach nicht beliebig gegen eins gehen.

Eine zweite Gruppe von Antworttechniken, die sich aus den Basisstrategien entwickelt hat, ist jene geschichteter Befragungsdesigns. Kim and Warde (2004) betrachten die Anwendung Warners Befragungstechnik in proportional bzw. optimal geschichteten Zufallsstichproben bei Ziehen mit Zurücklegen. Dabei wird in jeder einzelnen von K Schichten Warners Methode eingesetzt. Als Gesamtschätzer $\widehat{\pi}_A^{KW}$ für π_A ergibt sich die mit den relativen Schichtgrößen gewichtete Summe der Schätzer $\widehat{\pi}_{A,k}^W$ für die Schichtanteile $\pi_{A,k}$ $(k=1,2,...,K)$. Kim and Elam (2005) verknüpften die Theorie der geschichteten Zufallsauswahlen mit dem Befragungsdesign von Mangat and Singh (1990). Und auch deren Gesamtschätzer $\widehat{\pi}_A^{KE}$ für π_A ergibt sich aus der mit den relativen Schichtgrößen gewichteten Summe der Schätzer $\widehat{\pi}_{A,k}^{MS}$ für die Schichtanteile $\pi_{A,k}$. Die Varianzen von $\widehat{\pi}_A^{KW}$ und $\widehat{\pi}_A^{KE}$ für Ziehen ohne und Ziehen mit Zurücklegen folgen der Theorie der geschichteten Zufallsauswahlen (siehe dazu etwa: Quatember (2001), S. 49ff).

Als Beispiel für eine randomisierte Antworttechnik, die ein zusätzliches Designelement aufweist und sich nicht ausschließlich aus Warners bzw. Greenberg et al.'s Vorgehen ableitet, wird Mangat (1994) in unsere Untersuchung miteinbezogen. Seine zweistufige Strategie, deren konkreter Verlauf dem Interviewer wiederum verborgen bleiben muss, verlangt vom Respondenten jedenfalls eine „ja-Antwort", wenn er zur Gruppe U_A gehört. Wenn dies nicht der Fall ist, dann ist Warners Befragungsdesign anzuwenden. Der für π_A erwartungstreue Schätzer $\widehat{\pi}_A^{M2}$ ist hierbei (ebd., S.94):

$$\hat{\pi}_A^{M2} = \frac{\hat{\pi}_y - 1 + p}{p} \tag{12}$$

Für Ziehen mit Zurücklegen gibt Mangat (1994) die Varianz dieses Schätzers an (vgl. ebd., S.94):

$$V(\hat{\pi}_A^{M2}) = \frac{\pi_A \cdot (1-\pi_A)}{n} + \frac{(1-\pi_A)\cdot(1-p)}{n \cdot p} \tag{13}$$

Bei Ziehen ohne Zurücklegen lässt sich die Schätzervarianz darstellen als:

$$V(\hat{\pi}_A^{M2}) = \frac{\pi_A \cdot (1-\pi_A)}{n} \cdot \frac{N-n}{N-1} + \frac{(1-\pi_A)\cdot(1-p)}{n \cdot p} \tag{14}$$

Der **Beweis** von (14) befindet sich im Anhang. Die jeweils rechten Summanden der beiden Schätzervarianzen (13) und (14) entsprechen dem Streuungszuwachs, der durch das gewählte Befragungsdesign im Vergleich zu einer erfolgreich durchgeführten direkten Befragung entsteht.

3 Schlussbemerkungen

Randomisierte Antworttechniken bieten die Möglichkeit, durch Bewahrung der Privatsphäre der Respondenten Schätzungen für unbekannte Parameter sensitiver Merkmale zu ermöglichen, bei denen die Schätzung durch direkte Befragung wegen Antwortverweigerung und bewusste Falschbeantwortung beeinträchtigt wäre. Die durch ein cleveres Befragungsdesign gewährleistete Erhöhung der Privatsphäre wird durch einen Genauigkeitsverlust der Schätzung erkauft, der je nach Befragungsstrategie unterschiedlich hoch ausfallen kann.

Die Schätzervarianzen für Anteilschätzer sind auch bei Ziehen ohne Zurücklegen formal darstellbar. Allgemeine Aussagen über das Ausmaß der Erhöhung der Schätzervarianz im Vergleich verschiedener solcher Techniken sind jedoch wegen der Komplexität der (mehr oder weniger) frei wählbaren Designparameter nur bedingt möglich. Die Basisdesigns nach Warner (1965) und Greenberg et al. (1969) lassen sich durch Anwendung geschichteter Stichprobenverfahren effizienter gestalten. Mehrstufige Vorgehensweisen und ergänzende Designelemente können die Effizienz noch weiter erhöhen, so dass eine An- näherung an die Qualität der Schätzung an jene der direkten Befragung möglich erscheint. Diese Annäherung stößt jedoch dort an ihre Grenzen, wo durch die Festlegung der Designparameter der wesentliche Unterschied zur direkten Befragung, das ist die Kooperationsbereitschaft der Respondenten, verloren geht. Wo für die einzelnen Methoden diese Grenzen zu ziehen sind, lässt sich letztlich nur in der praktischen Umsetzung der Verfahren testen.

4 Anhang

4.1 Beweis von (5):

Es gilt:

$$V(\hat{\pi}_A^{G}) = \frac{1}{p^2} \cdot V(\hat{\pi}_y) \tag{15}$$

Es seien hinsichtlich der Erhebungseinheit i:

$$v_i = \begin{cases} 1 & \text{wenn i die Eigenschaft A aufweist,} \\ 0 & \text{sonst,} \end{cases}$$

$$w_i = \begin{cases} 1 & \text{wenn i die Eigenschaft B aufweist,} \\ 0 & \text{sonst,} \end{cases}$$

$$x_i = \begin{cases} 1 & \text{wenn i die Frage nach der Mitgliedschaft in } U_A \text{ erhält,} \\ 0 & \text{sonst,} \end{cases}$$

$$y_i = \begin{cases} 1 & \text{wenn i mit „ja“ antwortet,} \\ 0 & \text{sonst.} \end{cases}$$

Ferner gilt:

$$y_i = v_i \cdot x_i + w_i \cdot (1 - x_i) = v_i \cdot x_i + w_i - w_i \cdot x_i$$

Es ist $\widehat{\pi}_y = \frac{\sum_s y_i}{n}$ und somit wird (15) zu:

$$V(\hat{\pi}_A^G) = \frac{1}{n^2 \cdot p^2} \cdot V(\sum\nolimits_s y_i) \tag{16}$$

Nun ist:

$$V(\sum\nolimits_s y_i) = E(\sum\nolimits_s y_i^2) + E(\sum\nolimits_{s(i \neq j)} y_i \cdot y_j) - E^2(\sum\nolimits_s y_i) \tag{17}$$

Für Ziehen mit und ohne Zurücklegen gilt gleichermaßen:

$$E(\sum\nolimits_s y_i^2) = E(\sum\nolimits_s y_i) = n \cdot (p \cdot \pi_A + (1-p) \cdot \pi_B) \tag{18}$$

Für den 2. Summanden in (17) gilt:

$$\begin{aligned} E(\sum\nolimits_{s(i \neq j)} y_i \cdot y_j) \quad = \quad & n \cdot (n-1) \cdot E(v_i \cdot v_j \cdot x_i \cdot x_j + w_i \cdot v_j \cdot x_j - w_i \cdot x_i \cdot v_j \cdot x_j + \\ & + v_i \cdot x_i \cdot w_j + w_i \cdot w_j - w_i \cdot w_j \cdot x_i - v_i \cdot x_i \cdot w_j \cdot x_j - \\ & - w_i \cdot w_j \cdot x_j + w_i \cdot w_j \cdot x_i \cdot x_j) \end{aligned}$$

Es ist:

$$E(v_i \cdot v_j) = \begin{cases} \pi_A^2 & \text{bei Ziehen mit Zurücklegen,} \\ \frac{\pi_A \cdot (N\pi_A - 1)}{N-1} & \text{bei Ziehen ohne Zurücklegen} \end{cases}$$

und

$$E(w_i \cdot w_j) = \begin{cases} \pi_B^2 & \text{bei Ziehen mit Zurücklegen,} \\ \frac{\pi_B \cdot (N\pi_B - 1)}{N-1} & \text{bei Ziehen ohne Zurücklegen} \end{cases}$$

Somit ergibt sich für Ziehen mit Zurücklegen:

$$E(\sum\nolimits_{s(i \neq j)} y_i \cdot y_j) = n \cdot (n-1) \cdot (\pi_A^2 \cdot p^2 + 2 \cdot \pi_A \cdot \pi_B \cdot p - 2 \cdot \pi_A \cdot \pi_B \cdot p^2 + \pi_B^2 - 2 \cdot \pi_B^2 \cdot p + \pi_B^2 \cdot p^2) \tag{19}$$

und für Ziehen ohne Zurücklegen:

$$
\begin{aligned}
E(\sum\nolimits_{s(i\neq j)} y_i \cdot y_j) &= n\cdot(n-1)\cdot(\frac{\pi_A\cdot(N\pi_A-1)}{N-1}\cdot p^2 + 2\cdot\pi_A\cdot\pi_B\cdot p - \\
&\quad -2\cdot\pi_A\cdot\pi_B\cdot p^2 + \frac{\pi_B\cdot(N\pi_B-1)}{N-1} - 2\cdot\frac{\pi_B\cdot(N\pi_B-1)}{N-1}\cdot p + \\
&\quad +\frac{\pi_B\cdot(N\pi_B-1)}{N-1}\cdot p^2) \qquad (20)
\end{aligned}
$$

Der Subtrahend auf der rechten Seite von (17) ist bei Ziehen mit wie auch bei Ziehen ohne Zurücklegen folgendermaßen darstellbar:

$$
E^2(\sum\nolimits_s y_i) = n^2\cdot(\pi_A^2\cdot p^2 + 2\cdot\pi_A\cdot\pi_B\cdot p - 2\cdot\pi_A\cdot\pi_B\cdot p^2 + \pi_B^2 - 2\cdot\pi_B^2\cdot p + \pi_B^2\cdot p^2) \quad (21)
$$

Für Ziehen mit Zurücklegen ergibt sich für (16) aus (17), (18), (19) und (21) das Ergebnis (7).

Für Ziehen ohne Zurücklegen ergibt sich auf Grund der beim 2. Summanden (20) auftretenden Abweichung im Vergleich zu Ziehen mit Zurücklegen ausgehend von (16) folgende Varianzdarstellung:

$$
\begin{aligned}
V(\hat{\pi}_A^G) &= \frac{\pi_y\cdot(1-\pi_y)}{n\cdot p^2} - \frac{1}{n^2\cdot p^2}\cdot[\frac{n\cdot(n-1)}{N-1}\cdot p^2\cdot\pi_A\cdot(1-\pi_A) + \\
&\quad +\frac{n\cdot(n-1)}{N-1}\cdot\pi_B\cdot(1-\pi_B) - 2\cdot\frac{n\cdot(n-1)}{N-1}\cdot p\cdot\pi_B\cdot(1-\pi_B) + \\
&\quad +\frac{n\cdot(n-1)}{N-1}\cdot p^2\cdot\pi_B\cdot(1-\pi_B)]
\end{aligned}
$$

Daraus ergibt sich endlich (8):

$$
\begin{aligned}
V(\hat{\pi}_A^G) &= \frac{\pi_y\cdot(1-\pi_y)}{n\cdot p^2} - \frac{n-1}{n\cdot(N-1)}\cdot[\pi_A\cdot(1-\pi_A) + \pi_B\cdot(1-\pi_B)\cdot(\frac{1}{p^2} - \frac{2}{p} + 1)] \\
&= \frac{\pi_y\cdot(1-\pi_y)}{n\cdot p^2} - \frac{n-1}{n\cdot(N-1)}\cdot[\pi_A\cdot(1-\pi_A) + (\frac{1-p}{p})^2\cdot\pi_B\cdot(1-\pi_B)]
\end{aligned}
$$

4.2 Beweis von (14):

Es seien v_i und y_i wie in Abschnitt (3.1) definiert. Für x_i gilt:

$$
x_i = \begin{cases} 1 & \text{wenn i die Frage nach der Mitgliedschaft in } U_A \text{ erhält, gegeben } i \in U_A \\ 0 & \text{sonst.} \end{cases}
$$

Es gilt ferner:

$$
y_i = v_i + (1-v_i)\cdot(1-x_i) = 1 - x_i + v_i\cdot x_i
$$

Für den 1. Summanden in (17) ergibt sich:

$$
E(\sum\nolimits_s y_i) = n\cdot(\pi_A + (1-\pi_A)\cdot(1-p)) \qquad (22)
$$

Der 2. Summand in (17) wird zu:

$$
\begin{aligned}
E(\sum\nolimits_{s(i\neq j)} y_i\cdot y_j) &= n\cdot(n-1)\cdot E(1 - x_i + v_i\cdot x_i - x_j + x_i\cdot x_j - v_i\cdot x_i\cdot x_j + \\
&\quad +v_j\cdot x_j - v_j\cdot x_i\cdot x_j + v_i\cdot v_j\cdot x_i\cdot x_j)
\end{aligned}
$$

Dies ergibt für Ziehen mit Zurücklegen:

$$E(\sum_{s(i\neq j)} y_i \cdot y_j) = n \cdot (n-1) \cdot (1 - 2 \cdot p + 2 \cdot \pi_A \cdot p + p^2 - 2 \cdot \pi_A \cdot p^2 + \pi_A^2 \cdot p^2) \quad (23)$$

Für Ziehen ohne Zurücklegen ergibt sich hinsichtlich des 2. Summanden aus (17):

$$E(\sum_{s(i\neq j)} y_i \cdot y_j) = n \cdot (n-1) \cdot (1 - 2 \cdot p + 2 \cdot \pi_A \cdot p + p^2 - 2 \cdot \pi_A \cdot p^2 + \frac{\pi_A \cdot (N\pi_A - 1)}{N-1} \cdot p^2) \quad (24)$$

Der Subtrahend in (17) ist für das Befragungsdesign nach Mangat (1994) gegeben durch:

$$E^2(\sum_s y_i) = n^2 \cdot (1 + 2 \cdot \pi_A \cdot p - 2 \cdot p + \pi_A^2 \cdot p^2 - 2 \cdot \pi_A \cdot p^2 + p^2) \quad (25)$$

Für Ziehen mit Zurücklegen ergibt sich für (16) aus (17), (22), (23) und (25) das Ergebnis (13).

Für Ziehen ohne Zurücklegen ergibt sich auf Grund der beim 2. Summanden (24) auftretenden Abweichung im Vergleich zu Ziehen mit Zurücklegen ausgehend von (16) folgende Varianzdarstellung:

$$\begin{aligned} V(\hat{\pi}_A^{M2}) &= \frac{1}{n \cdot p^2} \cdot (n \cdot \frac{N-n}{N-1} \cdot p^2 \cdot \pi_A \cdot (1 - \pi_A) + n \cdot p - n \cdot p \cdot \pi_A - n \cdot p^2 + \\ &\quad + n \cdot p^2 \cdot \pi_A) \\ &= \frac{\pi_A \cdot (1 - \pi_A)}{n} \cdot \frac{N-n}{N-1} + \frac{(1 - \pi_A) \cdot (1 - p)}{n \cdot p} \end{aligned}$$

Und dies entspricht (14).

Literatur

Greenberg, B., A.-L. Abul-Ela, W. Simmons, and D. Horvitz (1969). The Unrelated Question Randomized Response Model: Theoretical Framework. *Journal of the American Statistical Association 64*, 520–539.

Horvitz, D., B. Shah, and W. Simmons (1967). The unrelated question randomized response model. *1967 Social Statistics Section Proceedings of the American Statistical Association*, 65–72.

Kim, J. and M. Elam (2005). A two-stage stratified Warner's randomized response model using optimal allocation. *Metrika 61*, 1–7.

Kim, J. and W. Warde (2004). A stratified Warner's randomized response model. *Journal of Statistical Planning and Inference 120*, 155–165.

Kim, J.-I. and J. Flueck (1978). Modifications of the Randomized Response Technique for Sampling Without Replacement. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 346–350.

Mangat, N. (1992). Two Stage Randomized Response Sampling Procedure Using Unrelated question. *Journal of the Indian Society of Agricultural Statistics 44*, 82–87.

Mangat, N. (1994). An Improved Randomized Response Strategy. *Journal of the Royal Statistical Society, Series B 56*, 93–95.

Mangat, N. and R. Singh (1990). An alternative randomized response procedure. *Biometrika 77*, 439–442.

Moors, J. (1971). Optimization of the Unrelated Randomized Response Model. *Journal of the American Statistical Association 66*, 627–629.

Quatember, A. (2001). *Die Quotenverfahren – Stichprobentheorie und -praxis*. Aachen: Shaker Verlag.

Warner, S. (1965). Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. *Journal of the American Statistical Association 60*, 63–69.

Kontaktadresse des Autors:
Ass. Prof. Dr. Andreas Quatember
Johannes Kepler Universität Linz
IFAS-Institut für Angewandte Statistik
Abteilung für Datengewinnung und Datenqualität
Altenbergerstraße 69
A-4040 Linz, Österreich
E-mail: andreas.quatember@jku.at